Preliminary Report on the Analysis of the Staten Island Pb Data

Introduction

The analysis aimed at

1. whether there was statistically significant evidence for lead from the site in the backyard samples at concentrations of, on average, greater than 400ppm,
2. or whether there was statistically significant evidence for concentrations, on average, less than 400ppm,
3. or, if not, what further sampling would ensure a high probability of confidently inferring that the average concentration is less than 400ppm were there no lead from the site in the backyard samples.

The data used were soil concentrations and isotope ratios from samples taken at the site, from samples taken off-site at locations assumed uncontaminated, and from samples taken in back-yards near the site. Only the "color-coded" observations were used. The analysis plan focused on examining to what extent the back-yard samples appeared to represent a mixture of lead characterized by the samples at the site and lead characterized by the samples at the uncontaminated locations.

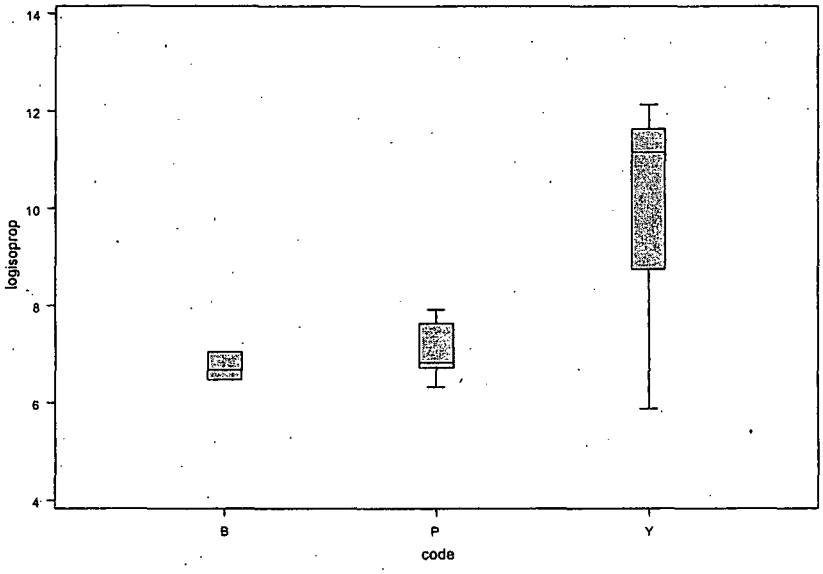
Data

The isotope ratio and concentration data used in the analysis are presented in Table 1. Concentrations of Pb206 and Pb208 together were computed as $C \times (r_1+r_2)/(1+ r_1+r_2)$, where C is the total measured concentration of lead, and r_1, r_2 are the 206/207 and 208/207 ratios. This calculation treats the presumably approximately 1% of Pb204 isotope as negligible. The values in the column labeled 'code' in the table represent the kind of sample, Y for onsite, P for back-yard, and B for uncontaminated.

r1	r2	ppm	code
2.393	1.122	90900	Y
2.390	1.117	240000	Y
2.397	1.126	147000	Y
2.400	1.125	8005	Y
2.411	1.138	456	Y
2.427	1.144	1250	P
2.439	1.160	3510	P
2.441	1.176	1020	B
2.451	1.180	1050	P
2.451	1.181	1110	P
2.453	1.179	841	B
2.454	1.184	1480	B
2.457	1.186	724	P
2.462	1.195	2620	P

The distribution of the log concentrations are depicted in Figure 1. It is evident from the figure that the concentrations in the site far exceed those in the other locations. But whether there is a difference between the back-ground and the back-yard samples is not obvious.

Figure 1.



Analysis

Let P denote the average concentration of Pb206 and Pb208 together off-site. Let Y denote the average concentration in backyards, and let q denote the proportion of PB206 and Pb208 in onsite lead. Let X denote the concentration of lead from onsite that is mixed in backyard samples. Then we have

$$P + qX - Y = 0.$$

We may estimate P using the average, in the uncontaminated sites, of the concentration of Pb206 and Pb208 together. We may estimate Y similarly from the back-yard sites. And we may estimate q from the concentrations in the onsite samples. The estimate of q may be adjusted for the presence of background lead, but because of the large concentrations onsite, the adjustment has minor impact on the results.

We may estimate q by solving the empirical version of the equation, and we may find a confidence interval for q by using the left hand side of the equation, normalized by its standard error, as a pivot statistic. Robust standard error computations that allow for different variances in the three samples were used in computing the standard errors that appear in the pivot statistic underlying the confidence interval.

Code for the analysis (in the SAS statistical package) is given below.

```
data data;
input name $ r1 r2 ppm code $;
datalines;
A-5-3 2.393 1.122 90900       Y
G-2-2 2.390 1.117 240000 Y
C-3-3 2.397 1.126 147000 Y
A-5-0 2.400 1.125 8005 Y
0-1 2.407 1.136 2760 0
B-2-0 2.411 1.138 456 Y
0-2 2.418 1.146 383    0
BY-034A 2.427 1.144 1250 P
TT-05A 2.436 1.168 396 0
BY-029A 2.439 1.160 3510 P
GP-38A 2.439 1.169 1070       0
GP-007B 2.441 1.176 1020 B
GP-008B 2.445 1.174 1330 0
TT-22A 2.448 1.179 2340       0
BY-13A 2.451 1.180 1050       P
BY-025A 2.451 1.181 1110 P
GP-006C 2.453 1.179 841       B
GP-006A 2.454 1.184 1480 B
BY-013C 2.457 1.186 724       P
GP-025A 2.461 1.196 1000 0
BY-025C 2.462 1.195 2620 P
;
run;
options mprint spool;
%macro m(start, by);
data data;
set data;
if code='0' then delete;
prop=(r1+r2)/(r1+r2+1);
P=0;
B=0;
Y=0;
if code='P' then P=1;
if code='B' then B=1;
if code='Y' then Y=1;
if code='Y' then z=prop;
if code='B' then z=prop*ppm;
if code='P' then z=prop*ppm;
run;
ods output acovtestanova=acovtestanova;
proc reg data=data;
model z= P B Y / noint hcc hccmethod=1;
%do i = &start %to 0 %by -&by;
testneg&i: test P-B+&i*Y = 0;
%end;
%do i = &by %to &start %by &by;
testpos&i: test P-B-&i*Y = 0;
%end;
run;
data acovtestanova;
set acovtestanova;
retain index -&start;
index=index+&by;
run;
%mend;
%m(2000, 10);
symbol1 color=black value=none interpol=join;
proc gplot data=acovtestanova;
plot probchisq*index;
run;
proc print data=acovtestanova;
where (abs(probchisq-0.05)<.0015) or (abs(probchisq-1.0)<0.008);
run;
```

Results

The value of X that solves the estimating equation is approximately 610ppm. The 90% two-sided confidence interval extends from 0ppm to 1580ppm; the two-sided p-value for testing that X=400 is 0.67, so that we may estimate the standard error of the estimate as (610-400)/0.44 = 477. Thus we cannot rule out that there is more than 400pm, pn average, contamination from the site, but there certainly is not statistically significant evidence that there is 400ppm or more.

In order to have power 0.9 to reject the one-sided null hypothesis that X exceeds 400ppm when the true value of X is zero at level 0.95, we set the standard error of X multiplied by 1.65+1.29 to 400 to obtain approximately 133. The ratio of 477 to 133 is approximately 3.5. Squaring 3.5, we arrive at requiring that the sample size be increased by a factor of 12. That is, there should be, instead of the 14 observations used, approximately 170 observations.